REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
31-10-2017	Final Report	9-Jun-2014 - 8-Jun-2017

4. TITLE AND SUBTITLE
Final Report: RESEARCH AREA 3: MATHEMATICS (3.1 Modeling of Complex Systems). Proposal should be directed to Dr. John Lavery

5a. CONTRACT NUMBER
W911NF-14-1-0254

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Universitat Politecnica De Valencia
Technology Transfer Office_CTT
UNIVERSITAT POLITÈCNICA DE VALÈNCIA
-

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
65349-MA.8

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
UU	UU	UU	UU		Paolo Rosso
					19b. TELEPHONE NUMBER
					963-877-007e

Standard Form 298 (Rev 8/98)
Prescribed by ANSI Std. Z39.18

RPPR Final Report
as of 29-Nov-2017

Agency Code:

Proposal Number: 65349MA **Agreement Number: W911NF-14-1-0254**

INVESTIGATOR(S):

Name: Paolo Rosso
Email: prosso@dsic.upv.es
Phone Number: 963877007ext73
Principal: Y

Organization: **Universitat Politecnica De Valencia**
Address: Technology Transfer Office_CTT, VALENCIA 46022,
Country: ESP

DUNS Number: 465011880 EIN:
Report Date: 08-Mar-2015 Date Received: 31-Oct-2017
Final Report for Period Beginning 09-Jun-2014 and Ending 08-Jun-2017
Title: RESEARCH AREA 3: MATHEMATICS (3.1 Modeling of Complex Systems). Proposal should be directed to Dr. John Lavery
Begin Performance Period: 09-Jun-2014 **End Performance Period:** 07-Aug-2017
Report Term: 0-Other
Submitted By: Elsa Cubel Email: ecubel@prhlt.upv.es
Phone: (349) 638-78170
Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: **STEM Participants:**

Major Goals: This project consisted of a twofold objective of detection of communities and copy in social networks, being the large scale of the available data the principal difficulty in its realization. As a result, the research to be conducted within this project was developed in two different directions: the detection of social circles and large-scale copy detection in textual documents.

Social circles represent communities among the users in a social network and, therefore, state-of-the-art community detection techniques can be applied for their detection. Our research within this problem was based on the application of several community detection methods, based on both clustering and deep-learning. We considered also the development of different novel feature representations for both the structure of the network itself and the users' profiles to feed the community detection algorithms. In addition, a critical discussion of the evaluation measures employed for the task in state-of-the-art works was needed, as they sometimes present serious flaws and lead to degenerate optimal performance.

The research on textual copy and similarity detection is focused on the application of different hashing techniques, with the aim to reduce the computational cost of both the detection of pairs of similar documents and the retrieval of similar documents to a given one. As most state-of-the-art hashing techniques are unsupervised, we incorporated supervision by learning metrics. Among the major goals of this research, the construction of a large-scale dataset of tweets about the topic ISIS was included. This dataset contains a portion of tweet pairs with similarity labels, obtained by means of a crowdsourcing process. We started to apply our methodology on publicly released Twitter datasets, as well. We considered several representations of tweets, from basic bags of words and n-grams to the novel, more complex, deep-learning based word embeddings.

Accomplishments: This report contains the work done during the project on both tasks of social circles detection and large-scale copy detection in Twitter.

Social circles detection

The core of our work was the development of several feature representations for both the network structure and the users' profiles. Structural features are based on the concepts of friendship ranks (the degree of closeness of the relationship between a certain pair of users), weighting (making the magnitude of the representation inversely proportional to the friendship rank that it represents) and aggregation (to represent all the friendship ranks with just one measure). With respect to profile features, we defined the following representations: explicit (they represent the

values of the profile characteristics taken by the users), intersection (they represent the relationship between pairs of users with respect to a certain profile characteristic) and weighted (they represent the relationship between pairs of users with respect to all the characteristics present in the experiments). In our study, we considered the 5 more informative profile characteristics: hometown, school, employers, gender and birthday; and we perform experiments using different subsets of them.
As a prediction technique, we employed multi-assignment clustering, a clustering method for vectorial data allowing for the assignment of the objects into more than one cluster, which allows to predict overlapping social circles. The evaluation of predicted sets of social circles is complex.
We followed two different approaches to define evaluation metrics. The first one consists in a similarity score between pairs of circles. To evaluate the whole set of predicted circles, an alignment between it and the set of groundtruth circles is needed. The alignments present in state-of-the-art studies present flaws and lead to degenerate optimal performance. The second evaluation framework is based on an edit distance between the set of groundtruth circles and the set of predicted circles. Our experiments were conducted by predicting a fixed number of circles and relying on the prediction technique to leave empty the extra circles. We compare our method to two basic baselines and a state-of-the-art method. Our method provides better results than the baselines and the state-of-the-art technique when they are evaluated by the edit distance measure, which is not the case when they are evaluated by the measures based on similarity scores between pairs of circles.
The results obtained from this research have led to the publication of two papers in peer-reviewed conference proceedings and the development of a master's thesis, while another paper has been published in the Neural Computing and Applications journal.
Alongside the extensive research on social circles detection, we constructed a dataset of text similarity between pairs of tweets, labelled by means of a crowdsourcing process.

Large-scale copy detection in Twitter

Within the framework of large-scale copy detection in Twitter, we depart from the similarity labelled dataset built by us during the project, along with the retrieval of a new dataset of 1 million unlabelled tweets, used to evaluate our method in large-scale scenarios. Our work was based on the learning of hash functions from the data. We hoped that similar pairs of tweets fell into the same hash code while dissimilar pairs of tweets fell into different hash codes, and also that the distribution of tweets into hash codes was uniform enough to reduce the search space of pairs of copied tweets. In our study we used different methods:
1) Large-Scale Copy Detection in Twitter Using Anchor Graph Hashing and Metric Learning
The objective of our study was to map documents into hash codes, so that
similar documents have to obtain the same hash code while dissimilar ones have to obtain different hash codes, in a pairwise manner. To evaluate this, we calculated the percentage of tweet pairs labelled as similar sharing the same hash code, and the percentage of tweet pairs labelled as dissimilar having different hash codes. Although both measures are important, we prioritised having a high value of the former one. Finally, we reported an accuracy measure consisting of an arithmetic mean of both measures. We conducted several hashing experiments, with and without Metric Learning, varying the tweet representations, the number of anchors and the number of bits of the hash codes. We observed that the hash functions obtained with Metric Learning are more accurate in mapping similar tweet pairs into the same hash code, which is our priority. In addition, generally, the experiments using Metric Learning are more accurate than the experiments without Metric Learning. We compared the performance of our methodology to two other state-of-the-art hashing methods: SimHash and Spectral Hashing . The performance of AGH in general, and AGH with Metric Learning in particular, is superior to the one of SimHash and Spectral Hashing. The quadratic cost of applying pairwise copy detection algorithms between every two documents in a certain textual dataset becomes unfeasible as the size of the dataset increases. Therefore, we applied hashing in order to search for copy detection just between pairs sharing the same hash code. We designed an empirical evaluation framework for the computational gain of our approach. This evaluation consisted in conducting hashing experiments in subsets of tweets of a size ranging from 50,000 to 1 million and comparing their cost to the one of checking every tweet pair in the dataset. The cost of our approach is much lower than the quadratic cost of the worst case. In addition, when the corpus size increases, the cost of the worst case grows faster than the cost of our approach that is empirically subquadratic. Therefore, we may conclude that it is assymptotically more efficient.

2) Large-Scale Copy Detection in Twitter Using Siamese Neural Networks
After obtaining the results using Anchor Graph Hashing and Metric Learning, we decided to conduct some textual similarity experiments using neural networks, which are nowadays state-of-the-art for a great variety of tasks. In particular, we chose a siamese architecture which is divided into two branches sharing weights. This way, the network can receive a pair of sentences as input and take the decision whether the sentences are similar or not. We have decided to adapt them for the task of textual similarity as we believe that they have potential to provide

good results. The architecture we used is composed of two siamese branches (sharing weights) which contain a word embeddings layer and several convolutional layers. After the concatenation of the output of both branches, we add optionally several fully connected layers, and finally we calculate a score from the output. The output score is calculated by means of the contrastive divergence, which is designed to pull together similar pairs and to push apart dissimilar pairs. We conducted several experiments both with and without fully connected layers at the end of the neural networks. We have also tried several values for the parameters. To evaluate the performance we calculate the accuracy, which is simply the percentage of tweet pairs that are correctly classified as similar or not similar.
The performance of these neural networks for the task is good and they can obtain a high accuracy, especially when adding fully connected layers at the end of the network. This work has resulted in the writing of a manuscript for the Journal of Engineering Business Management (awaiting revision).
In the Upload section, we include the report containing the work done on this task during the last twelve months of the project.

Training Opportunities: During the reporting period, one student has started his PhD studies.

Results Dissemination: Journal publications

- Jesús Alonso, Roberto Paredes and Paolo Rosso. Feature representation for social circles detection using MAC. Neural Computing and Applications, Volume 28, Issue 9, pp 2395–2402, September 2017. DOI: https://doi.org/10.1007/s00521-016-2222-y

- Jesús Alonso, Roberto Paredes and Poalo Rosso. Large scale copy detection with metric learning and hashing. International Journal of Engineering Business Management (submitted)

Conference publications

- Jesus Alonso, Roberto Paredes and Paolo Rosso. Data Mapping by Restricted Boltzmann Machinesfor Social Circles Detection. In: Proc. IEEE Int. Joint Conf. on Neural Networks (IJCNN), Killarney, Ireland, July 12-17. DOI: 10.1109/IJCNN.2015.7280653

- Jesús Alonso, Roberto Paredes, Paolo Rosso. Empirical Evaluation of Different Feature Representations for Social Circles Detection. In: Proc. 7th Iberian Conf. on Pattern Recognition and Image Analysis (ibPRIA), Pattern Recognition and Image Analysis, Springer-Verlag, LNCS(9117), pp. 31-38

Honors and Awards: Nothing to Report

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: Faculty
Participant: Paolo Rosso
Person Months Worked: 7.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Faculty
Participant: Enrique Vidal
Person Months Worked: 7.00 **Funding Support:**
Project Contribution:

International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Faculty
Participant: Roberto Paredes
Person Months Worked: 7.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: María Teresa Giménez
Person Months Worked: 5.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Jesús Alberto Alonso
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

CONFERENCE PAPERS:

Publication Type: Conference Paper or Presentation **Publication Status:** 4-Under Review
Conference Name: The Thirtieth Annual Conference on Neural Information Processing Systems (NIPS)
Date Received: Conference Date: 05-Dec-2016 Date Published: 31-Jul-2016
Conference Location: Barcelona (Spain)
Paper Title: Large-Scale Copy Detection with Metric Learning and Hashing
Authors: Jesús Alonso, Roberto Paredes, Paolo Rosso
Acknowledged Federal Support: **Y**

DISSERTATIONS:

RPPR Final Report
as of 29-Nov-2017

Publication Type: Thesis or Dissertation
Institution:
Date Received: 31-Jul-2015 Completion Date:
Title: Advances in Social Circles Detection
Authors:
Acknowledged Federal Support:

Final Report

1 Foreword

This report contains the work done by our team during the last twelve months. This work is focused on the task of large-scale copy detection in Twitter, based on the dataset about the topic ISIS built by us during previous reporting periods. Up to the moment, our work has resulted in the writing of a manuscript that is awaiting revision for the Information Processing Letters journal.

2 Large-Scale Copy Detection in Twitter Using Anchor Graph Hashing and Metric Learning

2.1 Introduction

Textual copy detection is among the most studied natural language processing problems, but recently it is gaining attention again, as copying has become a common phenomenon in the Web and social networks. The classical algorithms, focused on long and grammatically correct texts [6], are facing difficulties to adapt correctly to the short and normally grammatically incorrect documents present in social media. In this context, the development of novel techniques is needed.

Most modern copy detection methods are based on the retrieval of a set of documents similar to a given query [10, 9]. Our approach has a different aim, to detect pairs of documents with a high level of similarity. However, this would imply to check every possible pair in a given set of documents, which is not scalable to large datasets. To solve this problem, we propose a method to reduce the space of document pairs to be checked by means of a learnt hash function. The search space would then be reduced to the pairs of documents sharing the same hash code.

2.2 Anchor Graph Hashing

The hashing technique that we employ in our study is Anchor Graph hashing, based on Spectral hashing. The hash code optimization presented by Spectral hashing [7] is NP-hard, which is overcome by applying spectral relaxation and calculating real instead of binary codes. Having this assumption, the hash codes are calculated by building a neighbourhood graph and computing its eigenvectors. However, due to computational restrictions, the data needs to be assumed uniformly distributed.

The Anchor Graphs [5] are defined to avoid the last assumption. They allow to define the neighbourhood graph between the data points and a much smaller number of central points known as anchors. The hash codes can be then calculated as the eigenvectors of a smaller square matrix.

2.3 Metric Learning

Anchor Graph hashing is unsupervised, but we wanted to incorporate the similarity labels available for certain pairs of tweets in our dataset to learn the hash functions. We have included this supervision by learning a weighted Euclidean metric and using it to build the neighbourhood graph. We learn the weights by maximizing the significance of the difference between the similarity measures of similar and dissimilar pairs, as calculated by the Wilcoxon test of ranks [8], which is equivalent to maximizing the Area under the ROC Curve.

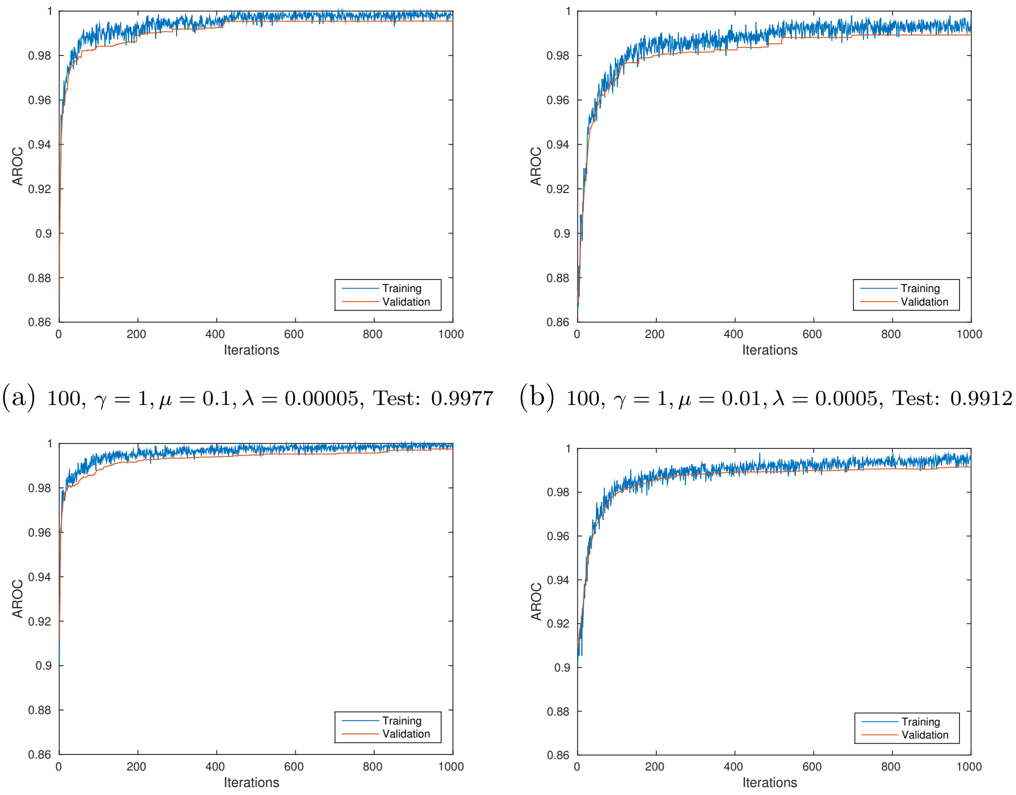
We employ a gradient descent methodology to optimize the weights, for which several parameters need to be adjusted. In Figure 1, some examples of the Metric Learning performance with different parameter configurations are shown. We have performed a grid search and selected two parameter configurations for our experiments.

2.4 Experiments

2.4.1 Dataset

We have used in our experiments the dataset retrieved from Twitter about the topic ISIS that we built during past reporting periods. This dataset is composed of pairs of tweets labelled as either similar or not similar. An example of a similar and a dissimilar pair can be seen in Table 1, and the size of the dataset is shown in Table 2.

Apart from this labelled dataset, we have collected an amount of 1 million tweets about ISIS to measure the computational cost of applying pairwise copy detection algorithms on large-scale datasets.



(a) 100, $\gamma = 1, \mu = 0.1, \lambda = 0.00005$, Test: 0.9977 (b) 100, $\gamma = 1, \mu = 0.01, \lambda = 0.0005$, Test: 0.9912

(c) 400, $\gamma = 1, \mu = 0.1, \lambda = 0.00005$, Test: 0.9964 (d) 400, $\gamma = 1, \mu = 0.01, \lambda = 0.0005$, Test: 0.9930

Figure 1: Metric Learning examples

In every case, the tweets have been preprocessed and represented as bags of word unigrams and bigrams of different sizes.

2.4.2 Similarity Detection

The objective of our study is to map documents into hash codes, so that similar documents have to obtain the same hash code while dissimilar ones have to obtain different hash codes, in a pairwise manner. To evaluate this, we calculate the percentage of tweet pairs labelled as similar sharing the same hash code, and the percentage of tweet pairs labelled as dissimilar having different hash codes. Although both measures are important, we prioritise having a high value of the former one. Finally, we report an accuracy measure consisting of an arithmetic mean of both measures. We have conducted several hashing experiments, with and without Metric Learning, varying the tweet representations, the number of anchors and the number of bits of the

Table 1: Example similar and not similar tweet pairs

	Tweet #1	Tweet #2
Similar pairs	ISIS Commander, Mullah Abdul Rauf, Killed In Afghanistan Drone Strike -	“No.2 Deputy head of #ISIS in #Afghanistan, Mullah Abdul Rauf, killed in air strike @ShababulAfghani”
Not similar pairs	#BobWoodward>#WhiteHouse Obsess DailyInterference #ISIS #War/#Pentagon wNO STRATEGY(Sabotage #military?) #independent	@enricof88 ISIS expands into Afghanistan, India keeps close watch. . . See —>

Table 2: Size of the annotated dataset, including the percentage of similar and dissimilar pairs of documents

	Number	Percentage
Total pairs of tweets	54,791	
Similar pairs	14,001	25.5%
Not similar pairs	40,790	74.5%

hash codes. The best results appear in Table 3.

We observe that the hash functions obtained with Metric Learning are more accurate in mapping similar tweet pairs into the same hash code, which is our priority. In addition, generally, the experiments using Metric Learning are more accurate than the experiments without Metric Learning.

2.4.3 Baselines

We have compared the performance of our methodology to two other state-of-the-art hashing methods: SimHash [2] and Spectral Hashing [7]. In Table 4 the best-performing results for each technique are presented, along with the best-performing AGH results, both with and without Metric Learning. The table shows that the performance of AGH in general, and AGH with Metric Learning in particular, is superior to the one of SimHash and Spectral Hashing.

Table 3: Best hashing results, with and without Metric Learning, for every document representation

	Without ML			With ML		
Representation	%SSC	%DDC	Accuracy	%SSC	%DDC	Accuracy
100	77.38%	92.62%	**85.00%**	99.19%	50.32%	74.75%
200	74.86%	93.63%	84.25%	98.74%	68.45%	83.60%
400	78.47%	83.12%	80.79%	98.88%	81.70%	90.29%
800	75.99%	87.57%	81.78%	98.90%	91.88%	**95.39%**
1600	62.88%	95.94%	79.41%	98.11%	92.01%	95.06%

Table 4: Best performing results, in comparison to other hashing methods

Method	Representation	%SSC	%DDC	Accuracy
SimHash	100	38.71%	98.45 %	68.58%
Spectral Hashing	100	39.79%	98.41%	69.10%
AGH without ML	100	77.38%	92.62%	85.00%
AGH with ML	800	98.90%	91.88%	**95.39%**

2.4.4 Computational Cost

The quadratic cost of applying pairwise copy detection algorithms between every two documents in a certain textual dataset becomes unfeasible as the size of the dataset increases. Therefore, we have applied hashing in order to search for copy detection just between pairs sharing the same hash code. We have designed an empirical evaluation framework for the computational gain of our approach. This evaluation consists in conducting hashing experiments in subsets of tweets of a size ranging from 50,000 to 1 million and comparing their cost to the one of checking every tweet pair in the dataset. The results are plotted in Figure 2.

The cost of our approach is much lower than the quadratic cost of the worst case. In addition, when the corpus size increases, the cost of the worst case grows faster than the cost of our approach that is empirically subquadratic. Therefore, we may conclude that it is assymptotically more efficient.

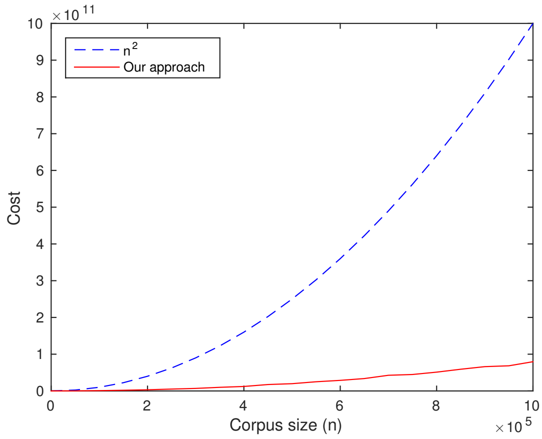


Figure 2: Computational cost of applying pairwise copy detection.

2.5 Publication

The results appearing in this section have been submitted in March 2017 to the Information Processing Letters journal as [1], which is awaiting revision.

3 Large-Scale Copy Detection in Twitter Using Siamese Neural Networks

3.1 Introduction

After obtaining the results mentioned in the former Section, we decided to conduct some textual similarity experiments using neural networks, which are nowadays state-of-the-art for a great variety of tasks. In particular, we chose a siamese architecture which is divided into two branches sharing weights. This way, the network can receive a pair of sentences as input and take the decision whether the sentences are similar or not.

3.2 Siamese Convolutional Neural Network Architecture

Siamese neural networks have been traditionally used for image classification [3], and are quite unexplored for textual tasks. We have decided to adapt them for the task of textual similarity as we believe that they have potential to provide good results. More concretely, the architecture that we have used is depicted in Figure 3. It is composed of two siamese branches (sharing weights) which contain a word embeddings layer and several convolutional layers. After the concatenation of the output of both branches, we add optionally several fully connected layers, and finally we calculate a score from the output.

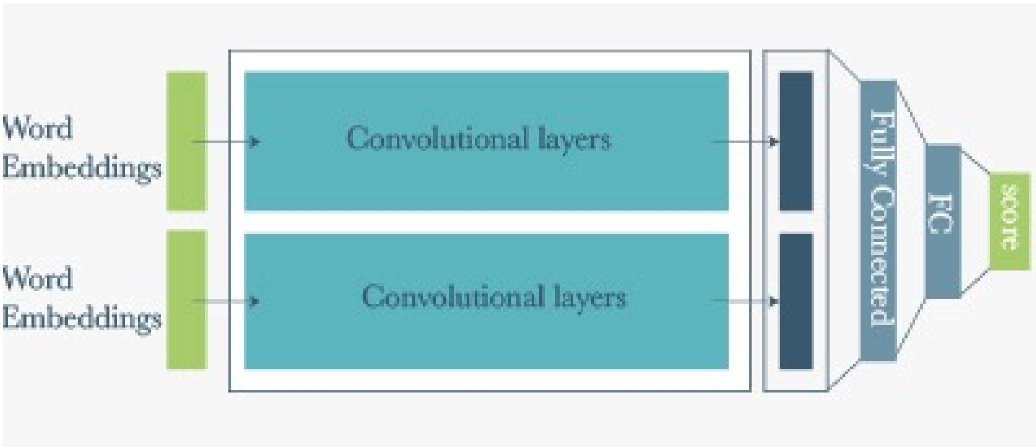


Figure 3: Siamese convolutional neural network architecture.

The output score is calculated by means of the contrastive divergence [4], which is designed to pull together similar pairs and to push apart dissimilar pairs.

3.3 Experiments

We have conducted several experiments both with and without fully connected layers at the end of the neural networks. We have also tried several values for the parameters. To evaluate the performance we calculate the accuracy, which is simply the percentage of tweet pairs that are correctly classified as similar or not similar.

The results can be seen in Table 5. The table shows that the performance of these neural networks for the task is good and they can obtain a high accuracy, especially when adding fully connected layers at the end of the network.

4 Bibliography

[1] J. Alonso, R. Paredes, and P. Rosso. Large-scale copy detection with metric learning and hashing. Submitted to the Information Processing Letters journal 2017.

Table 5: Results using Siamese Neural Networks

Architecture	Margin	Threshold	Accuracy
Siamese CNN	1.0	1.0	25.77%
Siamese CNN	1.0	0.5	78.37%
Siamese CNN	1.7	1.0	83.32%
Siamese CNN	1.5	1.0	94.81%
Siamese CNN + Fully (8)	1.5	1.0	90.95%
Siamese CNN + Fully (16)	1.5	1.0	96.30%
Siamese CNN + Fully (32)	1.5	1.0	**96.67%**
Siamese CNN + Fully (64)	1.5	1.0	96.40%
Siamese CNN + Fully (128)	1.5	1.0	94.66%

[2] Moses S Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388. ACM, 2002.

[3] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 539–546. IEEE, 2005.

[4] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, pages 1735–1742. IEEE, 2006.

[5] Wei Liu, Jun Wang, Sanjiv Kumar, and Shih-Fu Chang. Hashing with graphs. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1–8, 2011.

[6] M. Potthast, M. Hagen, A. Beyer, M. Busse, M. Tippmann, P. Rosso, and B. Stein. Overview of the 6th international competition on plagiarism detection. In L. Cappellato, N. Ferro, M. Halvey, and W. Kraaij, editors, *CLEF 2014 Labs and Workshops, Notebook Papers*, volume 1180, pages 951–957. CEUR-WS.org, 2014.

[7] Yair Weiss, Antonio Torralba, and Rob Fergus. Spectral hashing. In *Advances in Neural Information Processing Systems*, pages 1753–1760, 2009.

[8] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[9] Qi Zhang, Jihua Kang, Jin Qian, and Xuanjing Huang. Continuous word embeddings for detecting local text reuses at the semantic level. In *Proceedings of the 37th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 797–806. ACM, 2014.

[10] Qi Zhang, Yan Wu, Zhuoye Ding, and Xuanjing Huang. Learning hash codes for efficient content reuse detection. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 405–414. ACM, 2012.